Eur päisches Patentamt

Eur pean Patent Office

Office européen des brevets

Bescheinigung

Certificate

Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr. **Patent application No.** **Demande de brevet n°**

02025847.1

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

R C van Dijk

Europäisches Patentamt | European Patent Office | Office européen des brevets

Anmeldung Nr:
Application no.: 02025847.1
Demande no:

Anmeldetag:
Date of filing: 19.11.02
Date de dépôt:

Anmelder/Applicant(s)/Demandeur(s):

International Business Machines Corporation
New Orchard Road
Armonk, NY 10504
ETATS-UNIS D'AMERIQUE

Bezeichnung der Erfindung/Title of the invention/Titre de l'invention:
(Falls die Bezeichnung der Erfindung nicht angegeben ist, siehe Beschreibung.
If no title is shown please refer to the description.
Si aucun titre n'est indiqué se referer à la description.)

A method of clustering a set of records

In Anspruch genommene Prioriät(en) / Priority(ies) claimed /Priorité(s) revendiquée(s)
Staat/Tag/Aktenzeichen/State/Date/File no./Pays/Date/Numéro de dépôt:

Internationale Patentklassifikation/International Patent Classification/ Classification internationale des brevets:

G06F17/30

Am Anmeldetag benannte Vertragstaaten/Contracting states designated at date of filing/Etats contractants désignées lors du dépôt:

AT BE BG CH CY CZ DE DK EE ES FI FR GB GR IE IT LI LU MC NL PT SE SK TR

DESCRIPTION

A method of clustering a set of records

Field of the invention

The present invention relates to the field of data processing, and more particularly without limitation, to the field of data clustering.

Background and prior art

Clustering of data is a data processing task in which clusters are identified in a structured set of raw data. Typically the raw data consists of a large set of records each record having the same or a similar format. Each field in a record can take any of a number of logical, categorical, or numerical values. Data clustering aims to group such records into clusters such that records belonging to the same cluster have a high degree of similarity.

A variety of algorithms is known for data clustering. The K-means algorithm relies on the minimal sum of Euclidean distances to center of clusters taking into consideration the number of clusters. The Kohonen-algorithm is based on a neural net and also uses Euclidean distances. IBM's demographic algorithm relies on the sum of internal similarities minus the sum of external similarities as a clustering criterion. Those and other clustering criteria are utilized in an iterative process of finding clusters.

A common disadvantage of such prior art clustering methods is that they are computationally expensive and require a lot of computing power. This is especially true for very large data sets.

It is therefore an object of the present invention to provide an improved method of clustering which requires less computing power. It is a further object of the present invention to provide a corresponding computer program product and data processing system.

Summary of the invention

The underlying problem of the invention is solved basically by applying the features laid down in the respective independent claims. Preferred embodiments of the invention are given in the dependant claims.

In essence the present invention provides for a computationally inexpensive method for accurately clustering of data records containing structured raw data. Each of the data records contains a sequence of attribute values of corresponding attributes. For each of the attributes of the structured set of raw data contained in the records a characteristic value is calculated by evaluating the attribute values of that attribute across the data records. For each of the attribute values a deviation from the corresponding characteristic value is calculated. Next the attributes of each record are sorted based on the deviations to provide a sequence of attributes which is then used as a key for clustering.

In accordance with a preferred embodiment of the invention the mean value or the median value of the attribute values of a certain attribute across the data records is calculated to provide the characteristic value.

In accordance with a further preferred embodiment of the invention the deviation of an attribute value is calculated by determining the difference between the attribute value and the corresponding characteristic value. The difference should then be normalized, preferably by dividing by that characteristic value.

In accordance with a further preferred embodiment of the invention the attributes of a record are sorted using the corresponding deviations for the evaluation of a sorting criterion. For example the attributes with their corresponding deviations are sorted in ascending or descending order. Preferably the same sorting criterion is applied for all considered records.

In accordance with a further preferred embodiment of the invention the clustering is performed based on the keys provided by sorting the attributes of the records. A user may select a criterion of a given number of criteria for evaluation of the keys for clustering of the data records. For example all data records which have the same first m attributes are put into the same cluster considering or not considering the sign of the deviations.

In accordance with a further preferred embodiment of the invention, the clustering result is refined by searching of best matching keys in other clusters for the records of the smallest cluster. This way the records contained in the smallest cluster are distributed to other clusters such that the total number of clusters is reduced. For identification of other clusters for a record in the smallest cluster a distance measure such as a Euclids distance can be utilised.

In addition or as an alternative to Euclids distance gravitation can be used for reducing the number of clusters(http://www.ticam.utexas.edu/~zeyun/pick.htm).

The present invention is particularly advantageous in that it provides an efficient and computationally inexpensive way to analyse the characteristics of unknown data. It is a further particular advantage of the invention that performance of the clustering method only needs two passes over the data.

Brief description of the drawings

In the following preferred embodiments of the invention are described in greater detail by making reference to the drawings in which:

Figure 1 is illustrative of a flow chart of a preferred embodiment of a method of the invention,

Figure 2 is a block diagram of a preferred embodiment of a computer system of the invention.

Detailed description

Figure 1 shows a flow chart for performing a method of clustering of data records containing structured raw data. Given are n records $r_1,\ldots,r_n$ with k numeric attributes $a_1\ldots,a_k$, where $a_i(r_j)$ is the value of the i-th attribute of the j-th record. In step 100 a characteristic value is calculated for each of the attributes. For a given attribute this is done by calculating a projection of the attribute values of this attribute across the records.

For example, the mean value is calculated as a characteristic value for each one of the attributes: For each attribute a_l, $l = 1,\ldots,k$, calculate the mean value μ over all records

$$\mu(a_l) = \frac{1}{n}\sum_{i=1}^{n} a_l(r_i) \qquad (1)$$

Instead of the mean values the median values can be calculated. The median value is calculated by determining the difference between a maximum attribute value of a considered attribute and a minimum attribute value of the considered attribute over all records divided by two. Alternatively any other equivalent of a mean or median value can be calculated instead. By means of such mean values, median values or equivalent values characteristic values are provided for each one of the attributes.

In step 102 the deviations of each attribute value of a considered record from the corresponding characteristic value are determined. For example the deviation of an attribute value from its characteristic value can be performed by calculating the difference between the attribute value and its characteristic value. Preferably the difference is divided by the characteristic value.

In step 104 the deviations which have been obtained for each of the records are used as a basis for sorting the attributes of this record. For example the attributes are sorted in ascending or descending order of the deviations. This way a key comprising an ordered list of attributes and associated deviations is provided for each one of the records.

Preferably the steps 102 and 104 are carried out as follows:

1. Consider record r_i.

2. Consider attribute a_j.

3. Calculate the deviation $\hat{a}_j(r_i)$ of $a_j(r_i)$ from the respective mean of attribute a_j. This can be done by, but is not limited to

$$\hat{a}_j(r_i) = \frac{a_j(r_i) - \mu(a_j(r_i))}{\mu(a_j(r_i))} \qquad (2)$$

   or any other deviation formula.

4. Repeat this for all attributes $a_i, \ldots, a_k$ of the record r_i by applying steps 2 and 3.

5. Rank the deviations $|\hat{a}_1(r_i)|, \ldots, |\hat{a}_k(r_i)|$ from the largest to the smallest, holding $\hat{a}_{l_1}(r_i), \ldots, \hat{a}_{l_k}(r_i)$. This ranking shows which attributes deviate the most from the mean of all records. For example, since $\hat{a}_{I1}(r_i)$ has the largest deviation from the respective mean value $\mu(a_{l_1})$, this means

that record r_i is most significantly set apart from all other records by attribute a_{l_1}. The largest value shows the biggest deviation from the rest of the data, and hence that attribute is very characteristic.

In step 106 the records are clustered based on the keys.

One approach for performing the clustering based on the keys is to put records having identical keys into the same cluster. However this may result in a too large number of clusters.

It is therefore preferred to define a similarity criterion. When the keys of two records fulfil the similarity criterion the records are put into the same cluster.

Let $\hat{a}_{l_1}(r_i), \ldots, \hat{a}_{l_k}(r_i)$ be the ranking, i.e. the key, of record r_i and $\hat{a}_{l_1}(r_j), \ldots, \hat{a}_{l_k}(r_j)$ be the ranking of record r_j.

Some examples for preferred criteria are given in the following:

Criterion A: r_i and r_j belong to the same cluster if the first m attributes of the respective keys are identical and share the same sign. For example, if the three most significant attributes (m = 3) are considered, the ranking of record r_i is

$$(\hat{a}_7(r_i), \hat{a}_2(r_i), \hat{a}_3(r_i), \hat{a}_9(r_i), \ldots) = -1.17, 0.95, 0.87, 0.56, \ldots$$

and the ranking of r_j is

$$(\hat{a}_7(r_j), \hat{a}_2(r_j), \hat{a}_3(r_j), \hat{a}_1(r_j), \ldots) = -1.46, 1.09, 0.89, 0.88, \ldots.$$

The records r_i and r_j belong to the same cluster, as the first three attributes of the keys are identical as well as the signs of the values.

But if the ranking of r_k was $(\hat{a}_7(r_k),\hat{a}_2(r_k),\hat{a}_3(r_k),\ldots) = -1.46, -1.09, 0.89, 0.88, \ldots$, the r_i and r_k would belong to different sections, because the signs of the second most distinguishing attribute $\hat{a}_2$ had a different sign compared to the respective value of record r_i.

Criterion B: r_i and r_j belong to the same cluster if the first m attributes are identical. For example, considering the previous example, records r_i and r_k would belong to the same section, though the sign of the second most distinguishing attribute is different.

Criterion C: r_i and r_j belong to the same section if the same attributes appear on the first m positions with identical signs. This criterion ignores the order in which the attributes appear.

For example, if $m = 3$, r_i like before and the ranking of r_j is $\hat{a}_2(r_j),\hat{a}_3(r_j),(\hat{a}_7(r_j),\hat{a}_1(r_j),\ldots,) = 0.72, 0.68, -0.42, 0.37, \ldots$ a_2, a_3 and a_7 are identical and share the same signs.

This criterion can be varied with ignoring the signs.

The resulting clustering can be further refined by reducing the number of the clusters. For example it can be desirable to dissolve a cluster having a small size, i.e. having a small number of records. This can be done by means of the following iterative process:

1. Rank the clusters by size

2. Select the smallest cluster

3. For each record of the cluster, find the one of the larger clusters that matches most of the significant attributes. If more than one cluster has to be considered, either choose the largest of these clusters or use some kind of distance measure to find the nearest cluster.

4. Repeat until the desired number of clusters has been reached or if the similarity of records and clusters is too small.

Figure 2 shows a corresponding data processing system 200. Data processing system 200 has a database 202 for storing records of structured data. Each of the records has attribute values $a_1, \ldots, a_k$. Each of the records has an associated data field for storing a key for that record and a data field for storing a cluster identifier. Initially the key- and cluster data fields are empty.

Further data processing system 200 has a module 204 for calculating of characteristic values for each one of the attributes. The calculation of the characteristic values can be performed as explained with respect to step 100 of figure 1.

Further, data processing system 200 has module 206 for calculation of the deviations of the attribute values. This calculation can be performed in accordance with above equation (2).

Module 208 of the data processing system 200 is used for sorting of the attributes of the data records by applying a sorting criterion on the deviations of the corresponding attribute values. This way a ranking of the deviations can be obtained for each record. The sorting can be performed as explained with respect to step 104 of figure 1.

Further data processing system 200 has modules 210, 212 and 214 for application of the respective criteria A, B and C. The criteria A, B and C are described above with respect to figure 1.

Further there is a user interface 216. By means of the user interface 216 the tabular data contained in database 202 can be visualised. Further a user can select a subset of the records contained in the database 202 for performing a clustering operation. Before the data clustering is performed the user selects one of the pre-defined clustering criteria A, B or C. Alternatively the user can define a user specific clustering criterion.

After the user has selected the set of records of the database 202 on which the data clustering is to be performed and after a criterion for data clustering has been selected or specified, the data clustering is initiated.

Firstly, the module 204 is invoked to calculate the characteristic values of the attributes. Next the module 206 is invoked to calculate the deviations of the attribute values from their corresponding characteristic values. By means of module 208 the attributes are sorted to provide a key for each one of the selected records. Next the module for applying the selected criterion is invoked, i.e. module 210 for applying criterion A, module 212 for applying criterion B or module 214 for applying criterion C. Alternatively a user specified module is invoked to apply the user specified criterion. As a result of the application of the selected or specified criterion the selected records are clustered. Records which are put into the same cluster are assigned the same cluster identifier which is entered into the corresponding data field within database 202.

LIST OF REFERENCE NUMERALS

200	data processing system
202	database
204	module
206	module
208	module
210	module
212	module
214	module
216	user interface

CLAIMS

1. A method of clustering a set of records, each of the records having attribute values of a set of attributes, the method comprising the steps of:

   - for each attribute of the set of attributes: determining a characteristic value for that attribute based on the attribute values of that attribute,

   - for each attribute value: determining a deviation from the characteristic value of the corresponding attribute,

   - for each record: sorting of the attributes based on the deviations to provide a key,

   - clustering of the records based on the key.

2. The method of claim 1, whereby a mean value of the attribute values of that attribute is calculated as the characteristic value.

3. The method of claim 1 or 2, whereby a median value of the attribute values of that attribute is determined as the characteristic value.

4. The method of claim 1, 2 or 3, whereby the deviation is calculated based on a difference between that attribute value and the characteristic value of the corresponding attribute.

5. The method of any one of the preceding claims 1 to 4, whereby the deviation is calculated by calculating the difference between that attribute value and the characteristic value of the corresponding attribute, and by dividing the difference by the characteristic value of the corresponding attribute.

6. The methods of any one the preceding claims 1 to 5, whereby the absolute values of the deviations of the attributes are used as a sorting criterion.

7. The method of any one of the preceding claims 1 to 6, whereby a first one of the set of records having the first key and a second one of the set of records having a second key are put into the same cluster, if the first and the second keys have identical sub-sequences of a first length.

8. The method of any one of the preceding claims 1 to 7, a first one of the records of the set of records having the first key and a second record of the set of records having a second key are put into the same cluster, if the first and second keys contain identical sub-sequences of absolute values of the deviations.

9. The method of any one of the preceding claims 1 to 8, whereby a first record of the set of records having a first key and a second record of the set of records having a second key are put into the same cluster, if the first key has a first sub-sequence and the second key has a second sub-sequence, the first and second sub-sequences comprising the same set of attributes.

10. The method of any one of the preceding claims 1 to 9, whereby a first record of the set of records having a first key and a second record of the set of records having a second key are put into the same cluster, if the first key has a first sub-sequence and the second key has a second sub-sequence and if the first and second sub-sequences comprise the same attributes irrespective of a sign of the deviations of the attributes.

11. The method of any one of the preceding claims 1 to 10, further comprising:

- identifying a cluster having the smallest number of records,
- for each record of the identified cluster: searching another cluster having records with best matching keys.

12. The method of claim 11, whereby the length of the subsequences is reduced for finding a best match.

13. The method of claims 11 or 12, whereby a distance measure is used to find another cluster for a record of the identified cluster.

14. The method of claim 13, whereby the distance measure is Euclids distance.

15. The method of any one of claims 11 to 14, whereby gravitation is used for reducing the number of clusters.

16. A computer program product, such as a digital storage medium, comprising computer program means for performing a method in accordance with any one of the preceding claims 1 to 15.

17. A data processing system comprising processing means for performing a method in accordance with any one of the preceding claims 1 to 15.

ABSTRACT

A method of clustering a set of records

The invention relates to a method of clustering a set of records, each of the records having attribute values of a set of attributes, the method comprising the steps of:

- for each attribute of the set of attributes: determining a characteristic value for that attribute based on the attribute values of that attribute,
- for each attribute value: determining a deviation from the characteristic value of the corresponding attribute,
- for each record: sorting of the attributes based on the deviations to provide a key,
- clustering of the records based on the key.

(Figure 1)

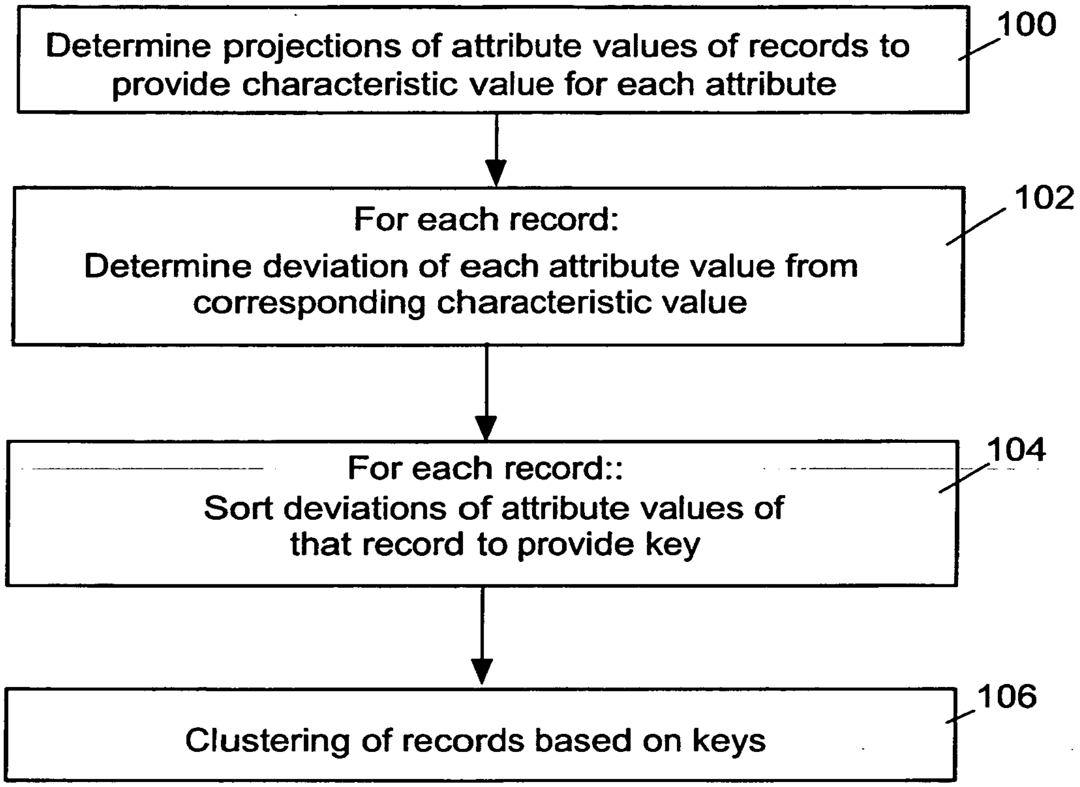
Determine projections of attribute values of records to provide characteristic value for each attribute
100
For each record:
Determine deviation of each attribute value from corresponding characteristic value
102
For each record::
Sort deviations of attribute values of that record to provide key
104
Clustering of records based on keys
106

FIG. 1

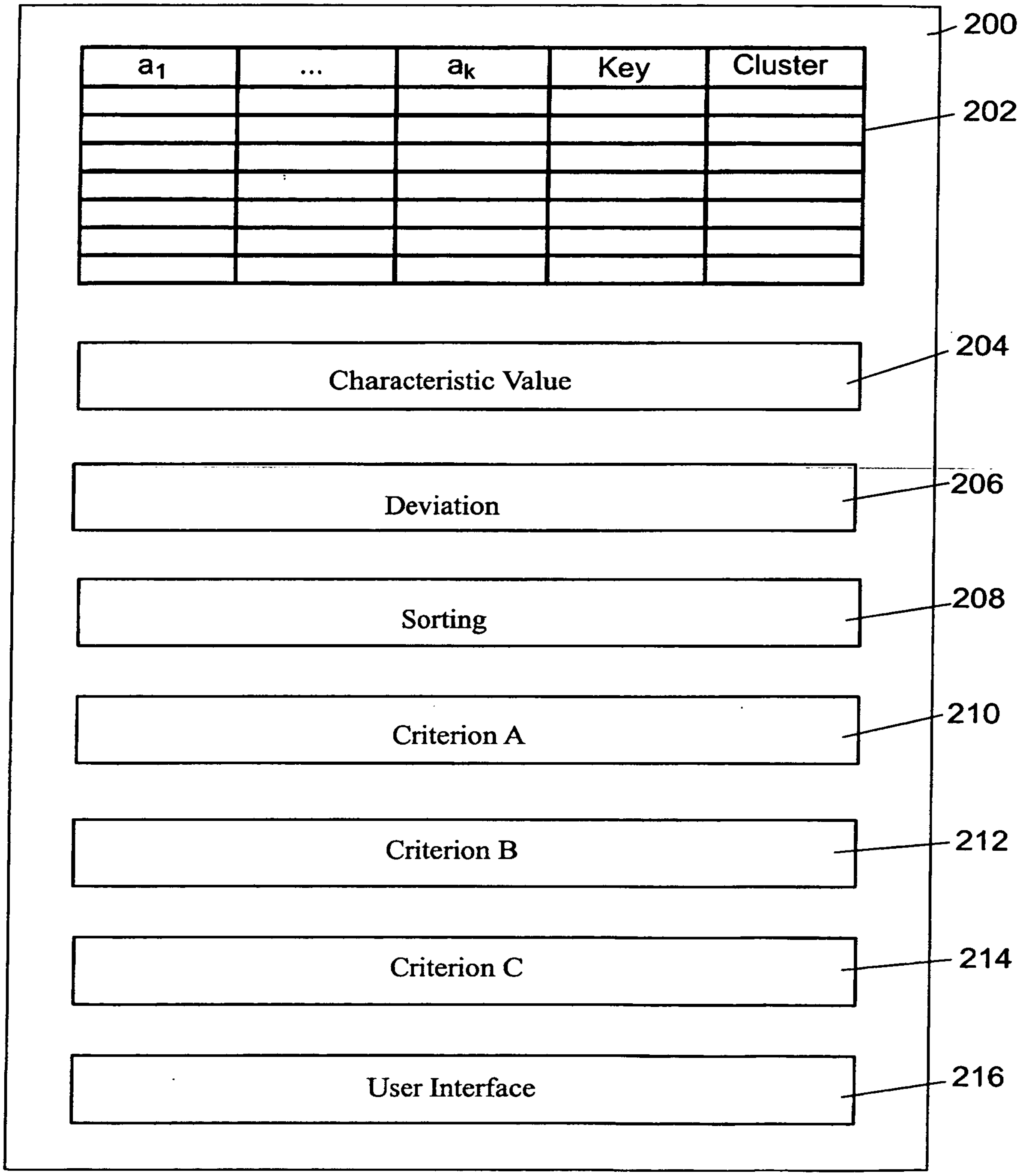
200
a_1
...
a_k
Key
Cluster
202
Characteristic Value
204
Deviation
206
Sorting
208
Criterion A
210
Criterion B
212
Criterion C
214
User Interface
216

FIG. 2